

### Exercise 1: High-dimensional Gaussian Distributions

Consider a random vector  $X = (X_1, \dots, X_p)^\top \sim \mathcal{N}(0, \mathbf{I})$ , i.e., a multivariate normally distributed vector with mean vector zero and covariance matrix being the identity matrix of dimension  $p \times p$ . In this case, the coordinates  $X_1, \dots, X_p$  are i.i.d. each with distribution  $\mathcal{N}(0, 1)$ .

- (a) Show that  $\mathbb{E}(\|X\|_2^2) = p$  and  $\text{Var}(\|X\|_2^2) = 2p$ , where  $\|\cdot\|_2$  is the Euclidean norm.

*Hint:*  $\mathbb{E}_{Y \sim \mathcal{N}(0,1)}(Y^4) = 3$ .

- (b) Use (a) to infer that  $|\mathbb{E}(\|X\|_2 - \sqrt{p})| \leq \frac{1}{\sqrt{p}}$  by using the following steps:

$$(i) \text{ Write } \|X\|_2 - \sqrt{p} = \underbrace{\frac{\|X\|_2^2 - p}{2\sqrt{p}}}_{=(1)} - \underbrace{\frac{(\|X\|_2^2 - p)^2}{2\sqrt{p}(\|X\|_2 + \sqrt{p})^2}}_{=(2)}.$$

- (ii) Compute  $\mathbb{E}[(1)]$ .

- (iii) Note that  $0 \leq \mathbb{E}[(2)] \leq \frac{\text{Var}(\|X\|_2^2)}{2p^{3/2}}$  holds due to  $\|X\|_2 \geq 0$ .

- (iv) Put (i)–(iii) together.

- (c) Use (b) to infer that  $\text{Var}(\|X\|_2) \leq 2$  by using the following steps:

- (i) Write  $\text{Var}(\|X\|_2) = \text{Var}(\|X\|_2 - \sqrt{p})$ .

- (ii) For any random variable  $Y$  it holds that  $\text{Var}(Y) \leq \mathbb{E}(Y^2)$ .

- (iii) If you encounter the term  $\mathbb{E}[\|X\|_2]$  write it as  $\mathbb{E}[\underbrace{\|X\|_2 - \sqrt{p}}_{= (*)} + \sqrt{p}]$  and use (b) for  $(*)$ .

- (d) Now let  $X' = (X'_1, \dots, X'_p)^\top \sim \mathcal{N}(0, \mathbf{I})$  be another multivariate normally distributed vector with mean vector zero and covariance matrix being the identity matrix of dimension  $p \times p$ . Further, assume that  $X$  and  $X'$  are independent, so that  $Z := \frac{X - X'}{\sqrt{2}} \sim \mathcal{N}(0, \mathbf{I})$ . Conclude from the previous that

$$|\mathbb{E}(\|X - X'\|_2 - \sqrt{2p})| \leq \sqrt{\frac{2}{p}} \quad \text{and} \quad \text{Var}(\|X - X'\|_2) \leq 4.$$

- (e) From the cosine rule we can infer that for any  $x, x' \in \mathbb{R}^p$  it holds that

$$\langle x, x' \rangle = \frac{1}{2}(\|x\|_2^2 + \|x'\|_2^2 - \|x - x'\|_2^2).$$

Use this to show that  $\mathbb{E}\langle X, X' \rangle = 0$ . Moreover, derive that  $\text{Var}(\langle X, X' \rangle) = p$ .

- (f) For different dimensions  $p$ , e.g.  $p \in \{1, 2, 4, 8, \dots, 1024\}$  create two sets consisting of 100 i.i.d. random observations from  $\mathcal{N}(0, \mathbf{I})$ , respectively and

- (i) compute the average Euclidean length of (one of) the sampled sets and compare it to  $\sqrt{p}$ ;
- (ii) compute the average Euclidean distances between the sampled sets and compare it to  $\sqrt{2p}$ ;
- (iii) compute the average inner products between the sampled sets;
- (iv) compute in (i)–(iii) also the empirical variances of the respective terms.

Visualize your results in an appropriate manner.